

|  |
| --- |
| CAPSTONE MODULE-2  D |
| Dr. Uzair Ahmad |
| October 263,2020  Credit Risk Analysis  Submitted by: Mohammed Imran (100805394)   Aman Verma (100799391) |

Contents

[1. Problem Statement 3](file:///C:\Users\aman1\Downloads\1103%20-%20Module%202.docx#_Toc54272707)

[2. Data Acquisition 3](file:///C:\Users\aman1\Downloads\1103%20-%20Module%202.docx#_Toc54272708)

[3. Exploratory Data Analysis 4](file:///C:\Users\aman1\Downloads\1103%20-%20Module%202.docx#_Toc54272711)

[4. Model Selection 8](file:///C:\Users\aman1\Downloads\1103%20-%20Module%202.docx#_Toc54272716)

[5. Summary](file:///C:\Users\aman1\Downloads\1103%20-%20Module%202.docx#_Toc54272716) 8

## Problem Statement

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.



Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience; Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

## Data Acquisition

The below dataset is being used to create a model which has been acquired from Kaggle.

2.1 Data Requirements

Application and previous history is required for model building. The labels are included in the training data and train a model to predict the labels from the features. Details of the data required are as follows

2.2 Data Preparation

Application Data {train test):

Static data for all applications. One row represents one loan in our data sample.

Bureau:

All client's previous credits history from credit

Transaction and Balances:

Customer credit transaction details from various channels like POS, cash Loans, and Credit cards.

Credit Card Balances:

Credit card monthly balances. Previous applications:

All previous applications for Credit loans of clients who have loans in our sample.

Repayment history:

For the previously disbursed credits in Credit related to the loans in our sample.

\*\*note that all of the above have multiple features in file.

## 3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an open-ended process where we calculate statistics and make figures to find trends, anomalies, patterns, or relationships within the data. The goal of EDA is to learn what our data can tell us. It generally starts out with a high-level overview, then narrows in to specific areas as we find intriguing areas of the data. The findings may be interesting in their own right, or they can be used to inform our modeling choices, such as by helping us decide which features to use.

## 3.1 Approach

The steps to be conducted for the EDA are: -

* To get an overview of the summary statistics of the data first need to retrieve the data.
* Examine the Distribution of the Target Column; The target is what we are asked to predict: either a 0 for the loan was repaid on time, or a 1 indicating the client had payment difficulties. We can first examine the number of loans falling into each category.
* Examine Missing Values; we can look at the number and percentage of missing values in each column. When it comes time to build our machine learning models, we will have to fill in these missing values (known as imputation).
* Column Types; Let's look at the number of columns of each data type; int64 and float64 are numeric variables (which can be either discrete or continuous). object columns contain strings and are categorical features.
* Encoding Categorical Variables; Before we go any further, we need to deal with pesky categorical variables. A machine learning model unfortunately cannot deal with categorical variables (except for some models such as LightGBM). Therefore, we have to find a way to encode (represent) these variables as numbers before handing them off to the model. There are two main ways to carry out this process:

## 3.2 Dataset – insights and cleaning

3.2.1 Numbers of feature in Train and Test dataset

app\_train.shape

(307511, 122)

app\_test.shape

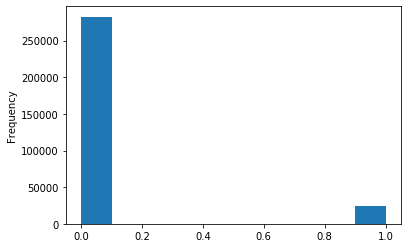
(48744, 121)

### 3.2.2 Target Column Distribution[¶](#Target-Column-Distribution)

0 282686

1 24825

Name: TARGET, dtype: int64



*Observation: Dataset is imbalance*

3.2.3 Missing – Null values

AMT\_REQ\_CREDIT\_BUREAU\_DAY 12.409732

AMT\_REQ\_CREDIT\_BUREAU\_WEEK 12.409732

AMT\_REQ\_CREDIT\_BUREAU\_MON 12.409732

AMT\_REQ\_CREDIT\_BUREAU\_QRT 12.409732

AMT\_REQ\_CREDIT\_BUREAU\_YEAR 12.409732

*Observations: 12 – 13 values are missing in some of the features in train and test set.*

3.2.4 Column Types

float64 65

int64 41

object 16

dtype: int64

*Observations: 16 column are object type , those have non number data set.*

3.2.5 Categorical feature and distinct classes in each feature

NAME\_CONTRACT\_TYPE 2

CODE\_GENDER 3

FLAG\_OWN\_CAR 2

FLAG\_OWN\_REALTY 2

NAME\_TYPE\_SUITE 7

NAME\_INCOME\_TYPE 8

NAME\_EDUCATION\_TYPE 5

NAME\_FAMILY\_STATUS 6

NAME\_HOUSING\_TYPE 6

OCCUPATION\_TYPE 18

WEEKDAY\_APPR\_PROCESS\_START 7

ORGANIZATION\_TYPE 58

FONDKAPREMONT\_MODE 4

HOUSETYPE\_MODE 3

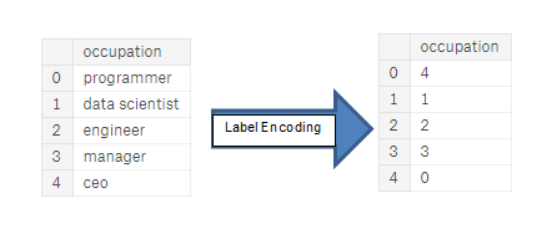
WALLSMATERIAL\_MODE 7

EMERGENCYSTATE\_MODE 2

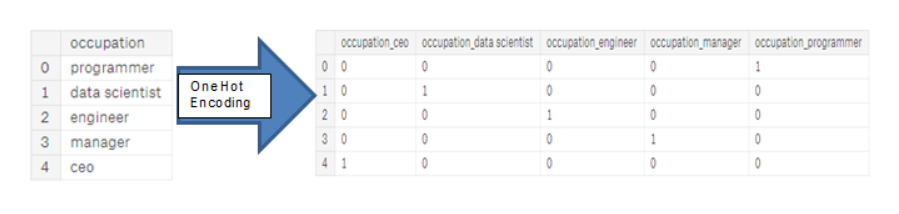
dtype: int64

## 3.3 Encoding

1. Label encoding: assign each unique category in a categorical variable with an integer. No new columns are created. An example is shown below



1. One-hot encoding: create a new column for each unique category in a categorical variable. Each observation receives a 1 in the column for its corresponding category and a 0 in all other new columns.



### Frequency Encoding : we will used frequency encoding as One hot is not suitable for Decision trees and Random Forest

## **3.**4 Anomalies

## 

### *Observations : Employed for 365243 days , too many year!!*[*¶*](#Employed--for-365243--days---,-too-many)

*Feature engineering the employment days , 365243 = 0*

count 252137.000000

mean -2384.169325

std 2338.360162

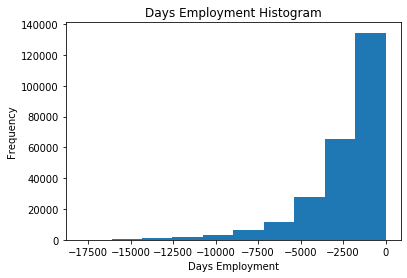
min -17912.000000

25% -3175.000000

50% -1648.000000

75% -767.000000

max 0.000000



\*\* - ve count backwards from current date

## 3.5 Feature Selection / Importance with Target

### 3.5.1 Correlations[¶](#Correlations)

Correlation give a general idea about target variable relationship with other features.

Most Positive Correlations:

OCCUPATION\_TYPE\_Laborers 0.043019

FLAG\_DOCUMENT\_3 0.044346

REG\_CITY\_NOT\_LIVE\_CITY 0.044395

FLAG\_EMP\_PHONE 0.045982

NAME\_EDUCATION\_TYPE\_Secondary / secondary special 0.049824

REG\_CITY\_NOT\_WORK\_CITY 0.050994

DAYS\_ID\_PUBLISH 0.051457

CODE\_GENDER\_M 0.054713

DAYS\_LAST\_PHONE\_CHANGE 0.055218

NAME\_INCOME\_TYPE\_Working 0.057481

REGION\_RATING\_CLIENT 0.058899

REGION\_RATING\_CLIENT\_W\_CITY 0.060893

DAYS\_EMPLOYED 0.074958

DAYS\_BIRTH 0.078239

TARGET 1.000000

Name: TARGET, dtype: float64

Most Negative Correlations:

EXT\_SOURCE\_3 -0.178919

EXT\_SOURCE\_2 -0.160472

EXT\_SOURCE\_1 -0.155317

NAME\_EDUCATION\_TYPE\_Higher education -0.056593

CODE\_GENDER\_F -0.054704

NAME\_INCOME\_TYPE\_Pensioner -0.046209

DAYS\_EMPLOYED\_ANOM -0.045987

ORGANIZATION\_TYPE\_XNA -0.045987

FLOORSMAX\_AVG -0.044003

FLOORSMAX\_MEDI -0.043768

FLOORSMAX\_MODE -0.043226

EMERGENCYSTATE\_MODE\_No -0.042201

HOUSETYPE\_MODE\_block of flats -0.040594

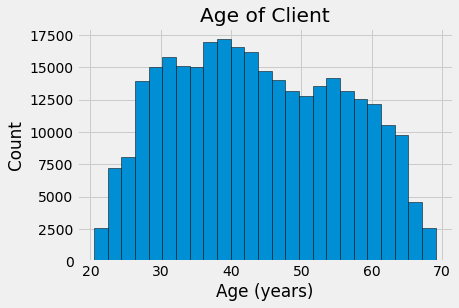
AMT\_GOODS\_PRICE -0.039645

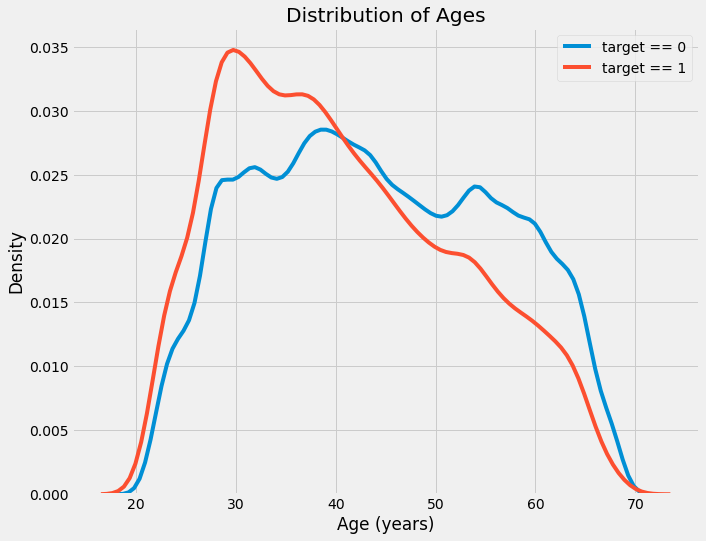
REGION\_POPULATION\_RELATIVE -0.037227

Name: TARGET, dtype: float64

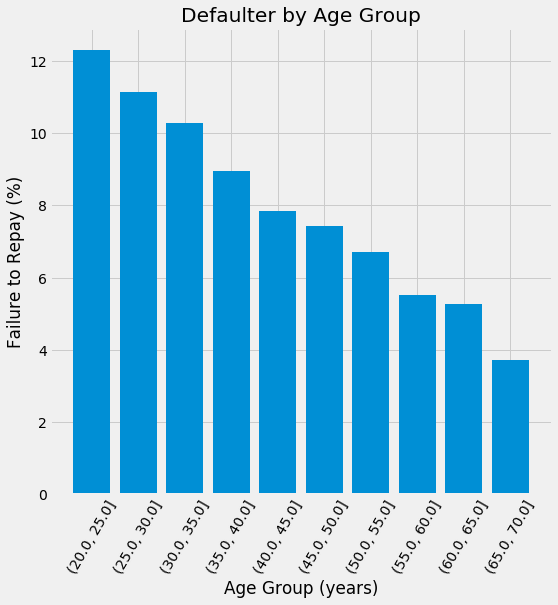
*Observation: there is no strong correlation between features and target*

3.5.2 Effect of Age feature on target repayment

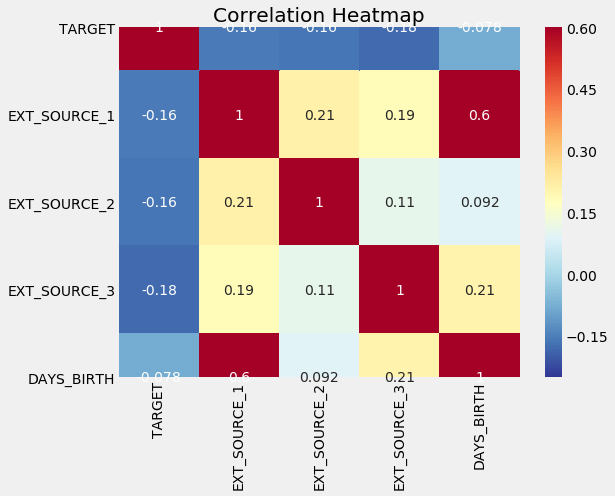


Age distribution in dataset

*Observation: Kernel density plot showing trend that younger are mote likely NOT to repay.*



*Observation: Bin – 5 years : showing same trend more clearly*



#### Observation : Moderate positive linear relationship between the EXT\_SOURCE\_1 and the DAYS\_BIRTH[¶](#Moderate-positive-linear-relationship-b)

3.5 Feature Engineering: Polynomial Features

3.5.1 Correlation between newly created polynomial feature and Target

EXT\_SOURCE\_2 EXT\_SOURCE\_3 -0.193939

EXT\_SOURCE\_1 EXT\_SOURCE\_2 EXT\_SOURCE\_3 -0.189605

EXT\_SOURCE\_2 EXT\_SOURCE\_3 DAYS\_BIRTH -0.181283

EXT\_SOURCE\_2^2 EXT\_SOURCE\_3 -0.176428

EXT\_SOURCE\_2 EXT\_SOURCE\_3^2 -0.172282

EXT\_SOURCE\_1 EXT\_SOURCE\_2 -0.166625

EXT\_SOURCE\_1 EXT\_SOURCE\_3 -0.164065

EXT\_SOURCE\_2 -0.160295

EXT\_SOURCE\_2 DAYS\_BIRTH -0.156873

EXT\_SOURCE\_1 EXT\_SOURCE\_2^2 -0.156867

Name: TARGET, dtype: float64

DAYS\_BIRTH -0.078239

DAYS\_BIRTH^2 -0.076672

DAYS\_BIRTH^3 -0.074273

TARGET 1.000000

1 NaN

Name: TARGET, dtype: float64

*Observation: no significant correlation found between polynomial feature and target*

4: Model Selection

* ****logistic regression baseline score around 0.671****
* ****Random Forest model score 0.678.****

# Light Gradient Boosting Machine model scores about 0.754

5: Summary

* *Dataset is highly imbalance, less then 10% of the values are 1*
* *ONE HOT ENCODING increased number of features, we would use \*\*\*count of frequency encoding*
* *Age is important as younger people are more likely to default*
* *Ext\_source\_1 feature is important as it has strong correlation with age*
* *Credit perios , Income and credit to income is also important.*
* *Number of days in employment has anomalies , its -ve and showing too many days for some rows.*
* *There are missing values found in some columns, will use imputation to balance data. Even polynomial feature engineering is not able to provide some strong correlation.*
* *Next step to merge all data sets and will find relationship " TARGET PERMUTATION "*